RUN PROBABILITIES IN SEQUENCES OF MARKOV-DEPENDENT TRIALS

Steven J. Schwager*

Biometrics Unit, Cornell University, Ithaca, New York 14853

BU-765-M Rev. August 1982

ABSTRACT

The probability of the occurrence of a run R is obtained as a function of the composition of R, the number n of trials, and the probabilities of the $v \geq 2$ possible outcomes at each trial. The run R can consist of any specified sequence of outcomes, and the probability that one or more of a given collection of runs occurs is also evaluated. The probabilities of the v possible outcomes can vary arbitrarily from trial to trial, and can be L-order Markov dependent on the L preceding outcomes. The practical application of these results is discussed.

KEY WORDS: Runs tests; Runs distributions; Multiple Markov dependence; Time-dependent trials.

AUTHOR'S FOOTNOTE

* Steven J. Schwager is Assistant Professor, Biometrics Unit, Cornell University, Ithaca, NY 14853. The author wishes to thank Barry Margolin, Steven Ostro, and Douglas Robson for helpful discussions about applications of the methods of this paper.

1. INTRODUCTION

Consider a series of n trials, $X_1, \cdots, X_n$, each of which has $v \geq 2$ possible outcomes, labeled 1, 2, ..., v for convenience. Define a run to be a specified sequence of outcomes that may occur at some point in the series of trials, e.g., $R_1 = (1,1,1,1,1) = 11111$, $R_2 = (1,1,2,2,1,1) = 112211$, and $R_3 = (1,2,4,1,2,4,1) = 1241241$. Runs containing a single symbol, e.g. R_1, will be called success runs. Runs containing at most two symbols, e.g. R_2, will be called success-failure runs. Runs containing arbitrarily many symbols, e.g. R_3, will be called multiple runs. The event that one or more of a given collection of runs occurs will be called a generalized run, e.g. $R_4 = R_1 \cup R_2 \cup R_3$.

This paper develops methods of computing the probability of the occurrence of a given success run, success-failure run, multiple run, or generalized run as a function of the composition of the run, the number n of trials, and the probabilities of the v possible outcomes at each trial. These probabilities will be assumed to depend on both the index t of the trial and the run e of length L immediately preceding trial t, where L is given. Let $p_{e,j;t}$ denote the probability that outcome j is observed at trial t given that run e has occurred in the preceding L trials. If $p_{e,j;t}$ is independent of t, it will be denoted by $p_{e,j}$, and the series of trials will be called L-order Markov dependent, where L is the length of e . When $L = 1$, the series of trials will be called simply Markov dependent. If $p_{e,j;t}$ is independent of the outcomes of all trials preceding trial t, it will be denoted by $p_{j;t}$, and the trials are independent. If $p_{e,j;t}$ is independent of both t and the outcomes of all preceding trials, it will be denoted by p_j, and the trials are IID (independent, identically distributed).

The probability of a specified run R will be obtained from recursive relations involving the probability f_m that R first occurs at trial m and the probability s_m that R has occurred at or before trial m . This recursion approach is superior to a generating function approach. The generating function $F(w) = \Sigma_{m=0}^{\infty} f_m w^m$ is defined from $f_m = P[\text{run R first occurs at trial m}]$ for $m \geq 1$, $f_0 = 0$. Feller (1968, Chapter XIII) used generating functions to treat success runs, success-failure runs, and generalized runs containing at most two symbols for the case of IID trials. Multiple runs and generalized multiple runs for IID trials can be handled by a straightforward extension of this discussion.

The generating function approach has several drawbacks. The greatest of these is that it applies only to trials with outcome probabilities independent of the trial index. Even in this case, the generating function method is not well suited to the exact calculation of P[run R occurs in n trials] $= \Sigma_{m=1}^{n} f_m$. This calculation utilizes a recursion involving f_m that is equivalent to the one developed below, but a greater amount of formal manipulation of the derivatives $F^{(m)}(0) = m!f_m$ is necessary to obtain it.

The rest of this paper is organized as follows. In Section 2, run probabilities for multiple runs under IID trials are derived. The assumption of identical trials is then removed, and generalized runs are treated. In Section 3, the assumption of independence of trials is replaced by first-order Markov dependence, and run probabilities are found. In Section 4, second-order Markov dependence is analyzed, and higher-order Markov dependence is discussed. In Section 5, generating functions for L-order Markov dependent trials are treated. In Section 6, applications of run probabilities and numerics are considered.

2. MULTIPLE RUNS UNDER INDEPENDENT TRIALS

Assume that the outcome probabilities at any trial are independent of

both the outcomes of all previous trials and the trial index t. The second assumption, that trials are identical, is not required for the approach of this section to be applicable, but it simplifies the presentation. When time dependence is introduced later in this section, the calculation becomes more complex but the method is essentially unchanged.

Let $T(m,m')$ with $1 \le m \le m' \le n$ denote the sequence of trial outcomes starting with trial m and ending with trial m', that is, $T(m,m') = (X_m, X_{m+1}, \cdots, X_{m'})$. For a multiple run R of length $k \ge 2$, let $R(i,i')$ with $1 \le i \le i' \le k$ denote the sequence of outcomes from entry i to entry i' of R, e.g. $R(1,k) = R$, and let $R(i) \equiv R(i,i)$. Two sequences are equal, e.g. $R = T(1,k)$, if they are equal entry by entry. Let B(m) denote the event that run R does not occur at or before trial m, where the occurrence of R at trial m means that $T(m-k+1,m) = R$. Run R occurs for the first time at trial m if R occurs at m and not at any $m' < m$.

This definition of the occurrence of R at trial m is simpler than that of Feller (1968, Chapter XIII); however, the analysis of this paper is unchanged if Feller's definition is substituted, as the discussion centers on the first occurrence of R, on which the two definitions agree. Consider the run R of length k occurring at trial m. The definition given above allows the inclusion of outcomes in $T(m-k+1,m) = R$ as part of another occurrence of R, so R can reoccur much earlier than trial $m+k$ if it has a repeating structure. In contrast, by Feller's definition, the listing of outcomes that may culminate in R begins again at trial $m+1$, so R cannot reoccur until trial $m+k$. For instance, in the sequence T = 11111111221122111, run R_1 of Section 1 occurs four times, at trials 5, 6, 7, and 8, and run R_2 of Section 1 occurs twice, at trials 12 and 16, by the definition of this paper; by Feller's definition, though, R_1 occurs only once, at trial 5, and R_2 occurs only once, at trial 12.

Let $p_1, \cdots, p_v$ denote the probabilities of outcomes 1, $\cdots$, v, respectively, and let $V \equiv \{1, 2, \cdots, v\}$. For $\ell \le k$, define $q(\ell) \equiv \prod_{i=k-\ell+1}^{k} p_{R(i)}$, so $q(\ell)$ is the probability $P[T(m-\ell+1,m) = R(k-\ell+1,k)]$ that the last ℓ entries of R occur on trials $m-\ell+1$ to m, where $m \ge \ell$. For example, for the run R_3 of Section 1, $R(1,4) = R(4,7) = 1241$, $R(3,7) = 41241$, $q(4) = p_1^2 p_2 p_4$, and $q(5) = p_1^2 p_2 p_4^2$. For $m \ge 1$, define

$$f_m \equiv P[\text{run R first occurs at trial m}] = P[B_{m-1}] - P[B_m] ,$$

$$s_m \equiv P[\text{run R occurs at or before trial m}] = 1 - P[B_m] .$$

Initial conditions are easy to establish:

$$f_m = s_m = 0 \qquad \text{for } m < k ; \tag{2.1}$$

$$f_m = s_m = q(k) \qquad \text{for } m = k . \tag{2.2}$$

Two recursive relations provide f_m and s_m for all $m > k$. The first is

$$s_m = s_{m-1} + f_m \qquad \text{for } m > k . \tag{2.3}$$

For the second, consider how R can occur for the first time at trial m. There can be no occurrence of R in T(1,m-k), and T(m-k+1,m) must equal R. Furthermore, the last few entries of T(1,m-k) cannot duplicate the first few entries of R in such fashion that when T(m-k+1,m) = R, the duplication will cause a completion of R before trial m. For instance, R_3 of Section 1 cannot occur for the first time at trial m if T(m-9,m-7) = (1,2,4), for T(m-6,m) = R_3 would imply that T(m-9,m-3) = R_3. In general, R cannot first occur at trial m if all of the following hold for some i:

1. $1 \le i \le k-1$;
2. $R(i+1,k) = R(1,k-i)$;
3. $T(m-k-i+1,m-k) = R(1,i)$.

Condition 1 makes i less than the full length of R, condition 2 indicates that the last k - i entries of R equal the first k - i entries of R, and condition 3 indicates that the last i trials before trial m - k + 1, where R should begin, equal the first i entries of R . These conditions may hold for more than one value of i simultaneously, as with R_3 of Section 1, where they hold for i = 3 whenever they hold for i = 6 .

Let I denote the set of all i satisfying conditions 1 and 2, and $I(m) \equiv \{i \in I : i \leq m-k\}$. For each $i \in I(m)$, define the event

$$C_{i,m} \equiv \{1 \text{ to } 3 \text{ hold for } i, \text{ and not for any larger } i' \in I(m)\} \quad .$$

Then

$$f_m = \{P[B_{m-k}] - \Sigma_{i \in I(m)} P[B_{m-k} \cap C_{i,m}]\} P[T(m-k+1,m) = R] \quad . \qquad (2.4)$$

Now observe that $P[T(m-k+1,m) = R] = q(k)$ and that when $B_{m-k} \cap C_{i,m}$ occurs, R can first occur at trial m-i but cannot possibly occur earlier. Therefore

$$\begin{aligned} P[B_{m-k} \cap C_{i,m}] &= P[B_{m-i-1} \cap T(m-k-i+1,m-k) = R(1,i)] \\ &= P[B_{m-i-1} \cap T(m-k-i+1,m-i) = R]/P[T(m-k+1,m-i) = R(i+1,k)] \\ &= f_{m-i}/q(k-i) \quad . \end{aligned}$$

Applying these results to (2.4) and noting that $i \in I - I(m)$ implies $m-i < k$, and hence $f_{m-i} = 0$, gives

$$f_m = \{1 - s_{m-k} - \Sigma_{i \in I} f_{m-i}/q(k-i)\} q(k) \text{ for } m > k \quad . \qquad (2.5)$$

This provides an expression for f_m in terms of s_{m-k} and $f_{m'}$ with $m' < m$. Initial conditions (2.1) and (2.2) and recursive relations (2.3) and (2.5) allow successive calculation of f_m and then s_m for each m from k to n .

Example: For R_1 (of Section 1), $k=5$, so $f_m = s_m = 0$ for $m < 5$ and $f_5 = s_5 = p^5$, where $p = p_1$. For $m > 5$, $I = \{1,2,3,4\}$ so $s_m = s_{m-1} + f_m$ and

$$f_m = [1 - s_{m-5} - \Sigma_{i=1}^{4} f_{m-i} p^{i-5}]p^5 \quad .$$

For R_2, $k=6$, so $f_m = s_m = 0$ for $m < 6$ and $f_6 = s_6 = p_1^4 p_2^2$. The set $I = \{4,5\}$, so for $m \geq 7$, $s_m = s_{m-1} + f_m$ and

$$f_m = [1 - s_{m-6} - f_{m-4}/p_1^2 - f_{m-5}/p_1]\, p_1^4 p_2^2 \quad .$$

For R_3, $k=7$, so $f_m = s_m = 0$ for $m < 7$ and $f_7 = s_7 = p_1^3 p_2^2 p_4^2$. The set $I = \{3,6\}$, so for $m \geq 8$, $s_m = s_{m-1} + f_m$ and

$$f_m = [1 - s_{m-7} - f_{m-3}/p_1^2 p_2 p_4 - f_{m-6}/p_1]p_1^3 p_2^2 p_4^2 \quad .$$

The same line of reasoning applies if the outcome probabilities $p_{1;t}$ to $p_{v;t}$ vary with the trial t. The only alterations necessary involve associating the appropriate trial index with the probabilities of the events considered. For $\ell \leq \min(m,k)$, define

$$q_m(\ell) \equiv \Pi_{j=k-\ell+1}^{k} p_{R(j);m-k+j} \quad .$$

Then

$$P[T(m-k+1,m) = R] = q_m(k) \quad ,$$

$$P[B_{m-k} \cap C_{i,m}] = f_{m-i}/q_{m-i}(k-i) \quad ,$$

and equations (2.2) and (2.5) are replaced by

$$f_m = s_m = q_k(k) \qquad \text{for } m = k \qquad (2.6)$$

$$= \{1 - s_{m-k} - \Sigma_{i \in I} f_{m-i}/q_{m-i}(k-i)\} q_m(k) \qquad \text{for } m > k \quad . \qquad (2.7)$$

The treatment just given is extended to the case of generalized runs with little difficulty. Only IID trials will be discussed, as the extension to the case of outcome probabilities varying with the trial t is

immediate. Let $R = R_1 \cup R_2 \cup \cdots \cup R_D$, that is, the generalized run R occurs when one or more of the multiple runs $R_1, \cdots, R_D$ occurs in a series of n trials. For $d = 1, \cdots, D$, let k_d denote the length of R_d, and define

$$q^d(\ell) \equiv \Pi_{j=k_d-\ell+1}^{k_d} p_{R_d}(j) = P[T(m-\ell+1,m) = R_d(k_d-\ell+1,k_d)] \quad \text{for } \ell \le k_d \ ,$$

$$f_m^d \equiv P[\text{run R first occurs at trial m and is of type d}] \ ,$$

$$f_m \equiv \Sigma_{d=1}^{D} f_m^d \ ,$$

and

$$s_m \equiv P[\text{run R occurs at or before trial m}] = \Sigma_{i=1}^{m} f_i \ .$$

The key idea needed for the analysis is that f_m^d can be found in exactly the manner just developed for f_m of a multiple run, except for the difference in the substituted value of s_{m-k_d} due to the possible occurrence of other components of R . Therefore

$$f_m^d = 0 \qquad \text{for } m < k_d \ ,$$

$$= q^d(k_d) \qquad \text{for } m = k_d \ ,$$

$$= \{1 - s_{m-k_d} - \Sigma_{i \in I_d} f_{m-i}^d / q^d(k_d - i)\} q^d(k_d) \qquad \text{for } m > k_d \ ,$$

$$f_m = \Sigma_{d=1}^{D} f_m^d \qquad \text{for } m \ge 1 \ ,$$

and

$$s_m = 0 \qquad \text{for } m = 1 \ ,$$

$$= s_{m-1} + f_m \qquad \text{for } m \ge 2 \ .$$

3. MARKOV DEPENDENT RUNS

Assume now that the outcome probabilities at any trial depend on the outcome of the previous trial, but not on the trial index t . As in the last section, the assumption that the trial index does not affect the outcome probabilities is made only to clarify the exposition, and will be dropped later.

Let $p_{h,j}$ denote the Markov transition probability of observing j at a particular trial given that the outcome h was observed at the previous trial. In this and all subsequent definitions, $h=0$ will be used for the initial trial, so $p_{0,j} = P[\text{trial 1 gives } j]$. Consider a specified multiple run R of length $k \geq 2$. Define for $\ell \leq k$

$$q(h;\ell) \equiv p_{h,R(k-\ell+1)} \prod_{i=k-\ell+2}^{k} p_{R(i-1),R(i)}$$

$$= P[T(m-\ell+1,m) = R(k-\ell+1,k) \,|\, T(m-\ell) = h] \quad .$$

Take f_m, s_m, and B_m as in the last section, and define for each $h \in V$ the event $B_{h;m} \equiv B_m \cap \{T(m) = h\}$, that R does not occur at or before trial m, on which outcome h is observed. Let $u_{h;m} \equiv P[B_{h;m}]$, so $s_m = 1 - \sum_{h=1}^{v} u_{h;m}$. A set of relations among the $v+1$ functions f_m, $u_{1;m}$, $\cdots$, $u_{v;m}$ will now be developed.

Initial conditions are again easy to establish:

$$f_m = 0 \qquad \text{for } m < k \quad ; \tag{3.1}$$

$$f_m = q(0;k) \qquad \text{for } m = k \quad ; \tag{3.2}$$

$$u_{j;m} = p_{0,j} \qquad \text{for all } j \in V,\ m = 1 \quad . \tag{3.3}$$

For the recursive relations, observe that the event $B_{j;m}$ occurs iff for some h, $B_{h;m-1}$ occurs, the mth trial then has outcome j, and R does not occur for the first time as a result. That is,

$$B_{j;m} = \cup_{h=1}^{v} B_{h;m-1} \cap \{T(m) = j\} \cap \{T(m-k+1,m) \neq R\} \quad ,$$

where the inequality involving R is automatically true when $j \neq R(k)$.

Thus

$$u_{j;m} = \sum_{h=1}^{v} u_{h;m-1} p_{h,j} \qquad \text{for } j \neq R(k),\ m \geq 2 \tag{3.4}$$

$$= \sum_{h=1}^{v} u_{h;m-1} p_{h,j} - f_m \qquad \text{for } j = R(k),\ m \geq 2 \quad . \tag{3.5}$$

A recursive formula for f_m can be found by incorporating Markov dependence into the derivation of (2.5):

$$\begin{aligned} f_m &= P[\cup_{h=1}^{v}\{B_{h;m-k} - \cup_{i\in I(m)} C_{i,m}\} \cap \{T(m-k+1,m) = R\}] \\ &= \Sigma_{h=1}^{v}\{P[B_{h;m-k}] - \Sigma_{i\in I(m)} P[B_{h;m-k} \cap C_{i,m}]\}q(h;k) \\ &= \Sigma_{h=1}^{v}\{u_{h;m-k} - \Sigma_{i\in I(m),\, R(i)=h} f_{m-i}/q(h;k-i)\}q(h;k) \\ &= \Sigma_{h=1}^{v} u_{h;m-k} q(h;k) - \Sigma_{i\in I} f_{m-i}\, q[R(i);k]/q[R(i);k-i] \qquad \text{for } m > k. \end{aligned} \tag{3.6}$$

When $u_{j;m'}$ and $f_{m'}$ are known for all j and $m' < m$, the value of f_m can be found from equation (3.6), after which the $u_{j;m}$ for all j follow from (3.4) and (3.5), and s_m is given by

$$s_m = 1 - \Sigma_{h=1}^{v} u_{h;m} \qquad \text{for all } m \ . \tag{3.7}$$

Example: For R_1 (of Section 1) with $v = 2$, these formulas provide probabilities of success runs of length 5 under Markov dependent trials, and are easily generalized to provide probabilities of success runs of arbitrary length k. Let $p_h \equiv p_{h,1}$ for $h = 0, 1, 2$, so $1 - p_h = p_{h,2}$. Then (3.1) to (3.3) give initial conditions $f_m = 0$ for $m \le 4$, $f_5 = p_0 p_1^4$, and $u_{1;1} = p_0$, $u_{2;1} = 1 - p_0$. Furthermore, $q(1;\ell) = p_1^{\ell}$ and $q(2;\ell) = p_2 p_1^{\ell-1}$, so

$$u_{1;m} = u_{1;m-1}p_1 + u_{2;m-1}p_2 - f_m \qquad \text{for } m \ge 2 \ , \tag{3.8}$$

$$u_{2;m} = u_{1;m-1}(1 - p_1) + u_{2;m-1}(1 - p_2) \qquad \text{for } m \ge 2 \ , \tag{3.9}$$

$$f_m = u_{1;m-5}p_1^5 + u_{2;m-5}p_2 p_1^4 - \Sigma_{i=1}^{4} f_{m-i} p_1^i \qquad \text{for } m \ge 6 \ , \tag{3.10}$$

and

$$s_m = 1 - u_{1;m} - u_{2;m} \qquad \text{for all } m \ . \tag{3.11}$$

For R_2 with $v = 2$, using the same notation, initial conditions are $f_m = 0$ for $m \le 5$, $f_6 = p_0 p_1^2 (1 - p_1) p_2 (1 - p_2)$, and $u_{1;1} = p_0$, $u_{2;1} = 1 - p_0$. Relations (3.8), (3.9), and (3.11) give $u_{1;m}$, $u_{2;m}$, and s_m for $m \ge 2$. The set $I = \{4, 5\}$, and $q(1;6) = p_1^3 (1 - p_1) p_2 (1 - p_2)$, $q(2;6) = p_1^2 (1 - p_1) p_2^2 (1 - p_2)$, $q(1;1) = p_1$, and $q(2;2) = p_1 p_2$, so for $m \ge 7$

$$f_m = u_{1;m-6}q(1;6) + u_{2;m-6}q(2;6) - [f_{m-4} + p_1 f_{m-5}]p_1(1-p_1)p_2(1-p_2) \quad .$$

For R_3 with $v=4$, (3.2) and (3.6) can be evaluated after noting that $q(h;7) = p_{h,1}p_{1,2}^2p_{2,4}^2p_{4,1}^2$, $q(4;4) = p_{1,2}p_{2,4}p_{4,1}^2$, and $q(4;1) = p_{4,1}$, giving

$$\begin{aligned} f_m &= q(0;7) && \text{for } m=7 \\ &= u_{1;m-7}q(1;7) + u_{2;m-7}q(2;7) + u_{3;m-7}q(3;7) + u_{4;m-7}q(4;7) \\ &\qquad - f_{m-3}q(4;7)/q(4;4) - f_{m-6}q(4;7)/q(4;1) && \text{for } m>7 \quad . \end{aligned}$$

The other steps in calculating s_m are routine.

If the outcome probabilities $p_{h,j;t}$ vary with the trial t, the only changes needed in the derivation of equations (3.1) to (3.7) are the insertion of appropriate trial indices in the probabilities. For $\ell \le \min(m,k)$, define $q_m(h;\ell) \equiv p_{h,R(k-\ell+1);m-\ell+1}\prod_{i=k-\ell+2}^{k} p_{R(i-1),R(i);m-k+i}$.

Then equations (3.1) to (3.6) are replaced by

$$\begin{aligned} f_m &= 0 && \text{for } m<k \\ &= q_k(0;k) && \text{for } m=k \\ &= \Sigma_{h=1}^{v} u_{h;m-k}q_m(h;k) - \Sigma_{i\in I} f_{m-i}q_m[R(i);k]/q_{m-i}[R(i);k-i] && \text{for } m>k \quad , \end{aligned}$$

$$\begin{aligned} u_{j;m} &= p_{0,j;1} && \text{for all } j\in V,\ m=1 \\ &= \Sigma_{h=1}^{v} u_{h;m-1}p_{h,j;m} && \text{for } j\neq R(k),\ m\ge 2 \\ &= \Sigma_{h=1}^{v} u_{h;m-1}p_{h,j;m} - f_m && \text{for } j=R(k),\ m\ge 2 \quad . \end{aligned}$$

Generalized runs can be analyzed by a slight modification of these formulas analogous to the approach resulting in the last three equations of Section 2.

4. HIGHER-ORDER MARKOV DEPENDENT RUNS

The results of the last section can be generalized to higher-order Markov dependent trials. This will be done explicitly for the second-order Markov model. Once more, the assumption that the outcome probabilities are independent of the trial index t is made temporarily to simplify the presentation.

Let $p_{gh,j}$ denote the second-order Markov transition probability of observing outcome j at a particular trial given that the outcome gh was observed in the previous two trials. Transition probabilities at trials 1 and 2 will be written with $g = h = 0$ and $g = 0$, respectively. For a specified multiple run R of length $k \geq 2$ and for $\ell \leq k$, define

$$q(gh;\ell) \equiv p_{gh,R(k-\ell+1)} p_{hR(k-\ell+1),R(k-\ell+2)} \prod_{i=k-\ell+3}^{k} p_{R(i-2,i-1),R(i)}$$
$$= P[T(m-\ell+1,m) = R(k-\ell+1,k) \,|\, T(m-\ell-1,m-\ell) = gh] \quad .$$

Take f_m, s_m, and B_m as usual, and define for each pair gh in $V \times V$ the event $B_{gh;m} \equiv B_m \cap \{T(m-1,m) = gh\}$, and its probability $u_{gh;m} \equiv P[B_{gh;m}]$, so $s_m = 1 - \sum_{g,h=1}^{v} u_{gh;m}$. Initial conditions are

$$f_m = 0 \qquad \text{for } m < k$$
$$f_m = q(00;k) \qquad \text{for } m = k$$
$$u_{hj;m} = p_{00,h} p_{0h,j} \qquad \text{for all } h, j \in V, \ m = 2 \quad .$$

Of course, if the length of R is $k = 2$, then $u_{R;m} = 0$ for all $m \geq 2$. The recursive relations among the v^2 functions $u_{hj;m}$ are given by

$$u_{hj;m} = \sum_{g=1}^{v} u_{gh;m-1} p_{gh,j} \qquad \text{for } hj \neq R(k-1,k), \ m \geq 3$$
$$= \sum_{g=1}^{v} u_{gh;m-1} p_{gh,j} - f_m \qquad \text{for } hj = R(k-1,k), \ m \geq 3 \quad .$$

The approach of earlier sections produces the equation

$$f_m = \Sigma_{g,h=1}^{v}\{P[B_{gh;m-k}] - \Sigma_{i\in I}P[B_{gh;m-k}\cap C_{i,m}]\}q(gh;k)$$
$$= \Sigma_{g,h=1}^{v}u_{gh;m-k}q(gh;k) - \Sigma_{g,h}\Sigma_{i\in I}P[B_{gh;m-k}\cap C_{i,m}]q(gh;k) \quad .$$

Reverse the order of summation in the double sum. For $i \in I$ with $i \geq 2$, the event $C_{i,m}$ is sufficient for $T(m-k-1,m-k) = R(i-1,i)$, so

$$\Sigma_{g,h=1}^{v}P[B_{gh;m-k}\cap C_{i,m}]q(gh;k) = P[B_{R(i-1,i);m-k}\cap C_{i,m}]q[R(i-1,i);k] \quad .$$

By condition 2 of Section 2, all symbols $R(j)$ for $j = 1, \cdots, k$ are the same, say r, iff $i = 1 \in I$. When $i = 1$, $C_{1,m}$ implies that $T(m-k-1) \neq T(m-k) = r$, as conditions 1 to 3 of Section 2 do not hold for i=2 when event $C_{1,m}$ occurs, so

$$\Sigma_{g,h=1}^{v}P[B_{gh;m-k}\cap C_{1,m}]q(gh;k) = \Sigma_{g\neq r}P[B_{gr;m-k}\cap C_{1,m}]q(gr;k)$$
$$= \Sigma_{g\neq R(1)}u_{gR(1);m-k}q[gR(1);k] \quad .$$

Combining these gives for $m = k+1$

$$f_m = \Sigma_{h=1}^{v}u_{Oh;m-k}q(Oh;k) - \Sigma_{i\in I,\, i=1}u_{OR(1);m-k}q[OR(1);k] \quad , \qquad (4.1)$$

and for $m > k+1$

$$f_m = \Sigma_{g,h=1}^{v}u_{gh;m-k}q(gh;k) - \Sigma_{i\in I,\, i\geq 2}f_{m-i}\,q[R(i-1,i);k]/q[R(i-1,i);k-i]$$
$$- \Sigma_{i\in I,\, i=1}\Sigma_{g\neq R(1)}u_{gR(1);m-k}q[gR(1);k] \quad . \qquad (4.2)$$

Finally, the probability $s_m = 1 - \Sigma_{g,h=1}^{v}u_{gh;m}$.

Markov dependence of order $L > 2$ can be handled by the same methods and principles used for $L = 1$ and 2 . Extra care is needed when $L > k$. Time dependent transition probabilities $p_{gh,j;t}$ and generalized runs can be analyzed by the same methods used in previous sections.

Example: For R_1 (of Section 1) with $v = 2$, let $p_{gh} \equiv p_{gh,1}$ for $g,h \in \{0,1,2\}$. Initial conditions are $f_m = 0$ for $m \leq 4$, $f_5 = p_{00}p_{01}p_{11}^3$, and $u_{11;2} = p_{00}p_{01}$,

$u_{12;2} = p_{00}(1 - p_{01})$, $u_{21;2} = (1 - p_{00})p_{02}$, $u_{22;2} = (1 - p_{00})(1 - p_{02})$. The recursive relations for the u's are, for $m \geq 3$,

$$u_{11;m} = u_{11;m-1}p_{11} + u_{21;m-1}p_{21} - f_m$$
$$u_{21;m} = u_{12;m-1}p_{12} + u_{22;m-1}p_{22}$$
$$u_{12;m} = u_{11;m-1}(1 - p_{11}) + u_{21;m-1}(1 - p_{21})$$
$$u_{22;m} = u_{12;m-1}(1 - p_{12}) + u_{22;m-1}(1 - p_{22}) \quad .$$

For f_m, recall that $I = \{1,2,3,4\}$, so

$$f_m = u_{02;m-5}q(02;5) = (1 - p_{00})p_{02}p_{21}p_{11}^3 \qquad \text{for } m = k+1 = 6 \quad ;$$
$$f_m = u_{12;m-5}q(12;5) + u_{22;m-5}q(22;5)$$
$$= (u_{12;m-5}p_{12} + u_{22;m-5}p_{22})p_{21}p_{11}^3 \qquad \text{for } m > 6 \quad ,$$

after using the probabilistic identity

$$u_{11;m-5}q(11;5) = \Sigma_{i=2}^{4} f_{m-i}p_{11}^i \quad .$$

The results of this example can be adjusted in a straightforward way to give probabilities of success runs of length k under second-order Markov dependent trials.

5. L-ORDER MARKOV GENERATING FUNCTIONS

Generating functions are not a helpful tool for calculating f_m, as mentioned at the end of Section 1; however, they are useful in examining large-sample behavior of the processes treated here. Generating functions for these processes will now be considered. Attention is restricted to the case of outcome probabilities $p_{e,j}$ that do not vary with the trial t . Markov dependence of any order L and generalized runs can be handled by the method of this section, but for ease of presentation the discussion deals only with Markov dependent ($L = 1$) multiple runs.

Take $f_0 = s_0 = 0$ and $u_{j;0} = 0$ for each $j \in V$, and recall that $f_m = s_m = 0$ for $m < k$ by (3.1). Define the generating functions

$$F(w) \equiv \Sigma_{m=0}^{\infty} f_m w^m = f_k w^k + f_{k+1} w^{k+1} + \cdots ,$$

$$S(w) \equiv \Sigma_{m=0}^{\infty} s_m w^m = s_k w^k + s_{k+1} w^{k+1} + \cdots ,$$

and

$$U_j(w) \equiv \Sigma_{m=0}^{\infty} u_{j;m} w^m \qquad \text{for } j = 1, \cdots, v .$$

Multiplying (3.4) by w^m, summing over $m = 2, 3, \cdots$, and adding (3.3) multiplied by w gives

$$U_j(w) = p_{0,j} w + \Sigma_{h=1}^{v} p_{h,j} w U_h(w) \qquad \text{for } j \neq R(k) . \tag{5.1}$$

The same procedure with (3.5) replacing (3.4) yields

$$U_j(w) = p_{0,j} w + \Sigma_{h=1}^{v} p_{h,j} w U_h(w) - F(w) \qquad \text{for } j = R(k) . \tag{5.2}$$

Multiplying (3.6) by w^m, summing over $m = k+1, k+2, \cdots$, and adding (3.2) multiplied by w^k produces

$$F(w) = q(0;k) w^k + \Sigma_{h=1}^{v} q(h;k) w^k U_h(w)$$
$$- (\Sigma_{i \in I} q[R(i);k] w^i / q[R(i);k-i]) F(w) . \tag{5.3}$$

Equations (5.1), (5.2), and (5.3) constitute a set of $v+1$ linear equations in $F(w)$, $U_1(w)$, $\cdots$, $U_v(w)$. Standard techniques can be applied to solve this set of equations. The general solution is quite complex, but the structure of a particular situation can lead to some simplification.

To obtain S(w), multiply (3.7) by w^m and sum over $m \geq 1$, giving

$$S(w) = w/(1-w) - \Sigma_{h=1}^{v} U_h(w) .$$

Alternatively, multiply the equation $s_m = s_{m-1} + f_m$ by w^m and sum over $m \geq k$,

resulting in $S(w) = (1 - w)^{-1}F(w)$.

Example: For R_1 (of Section 1) with $v = 2$ and the notation from Section 3, (5.2), (5.1), and (5.3) yield

$$F(w) + (1 - p_1 w)U_1(w) - p_2 w U_2(w) = p_0 w \quad ,$$

$$-(1 - p_1)wU_1(w) + [1 - (1 - p_2)w]U_2(w) = (1 - p_0)w \quad ,$$

$$\left[\Sigma_{i=0}^{4} p_1^i w^i\right] F(w) - p_1^5 w^5 U_1(w) - p_2 p_1^4 w^5 U_2(w) = p_0 p_1^4 w^5 \quad .$$

These three equations can be solved by Cramer's rule or other approaches, giving after some algebra

$$F(w) = p_1^4[p_0 + (p_2 - p_0)w][1 - p_1 w]w^5/G \quad ,$$

$$U_1(w) = [1 - p_1^4 w^4][p_0 + (p_2 - p_0)w]w/G \quad ,$$

$$U_2(w) = [(1 - p_0)w + (p_0 - p_1)w^2 - p_0 p_1^4(1 - p_1)w^6]/G \quad ,$$

where

$$G = 1 - (p_1 + 1 - p_2)w + (p_1 - p_2)w^2 + p_1^4(1 - p_1)p_2 w^6 \quad .$$

The generating function $S(w)$ is obtained immediately either from $F(w)$ or from $U_1(w)$ and $U_2(w)$. These formulas are generalized to success runs of arbitrary length k without difficulty.

6. APPLICATIONS AND NUMERICS

Outcome probabilities exhibiting Markov dependence and variation over trials appear in models from numerous fields of application. A few of these will now be mentioned.

DNA sequencing. A molecule of deoxyribonucleic acid is a chain or sequence of nucleotides with the four base structures adenine, cytosine, guanine, and thymine, or A, C, G, and T, which form three-letter genetic "words". The type of nucleotide occupying a given position in the chain

can be modeled as being dependent on the types of the several nucleotides immediately preceding it. The occurrence of a specified sequence of nucleotides in some portion of the chain is the event that the specified run of A's, C's, G's, and T's occurs.

Psychology. The theory that success breeds success, i.e., that attaining a positive outcome makes it more probable that a positive outcome will be attained on the next trial, is often considered in psychological achievement testing, animal learning studies, athletic competition, and similar matters. It is also possible that failure breeds failure, or both phenomena may be present. The probabilities of success runs in several different situations of this general type are calculated at the end of this section.

In animal learning experiments, a test animal performs a sequence of trials, in each of which it either succeeds or fails at selecting a box containing food, running a maze, or some other task. The length of the longest run is often used as a test for improvement in performance (Bradley 1968), which occurs when $p_{1;t}$, the probability of success at trial t, increases with t . The run probabilities in Table 1 below make it clear that great care must be taken in concluding that learning is taking place, as Markov dependence with fixed transition probabilities also tends to produce longer runs than IID trials.

Sociology. The behavior of groups of people forming lines and other structures can be modeled as a Markov dependent sequence of trials in which some characteristic, such as the sex of the individual, is taken as the trial outcome. It is of interest to know whether the occurrence of certain runs is plausible under various types and orders of Markov dependence.

Ecology. When a line or strip transect is taken in a wooded area, certain characteristics of the sampled trees intersecting the transect are recorded, e.g., species, presence of disease, bark cover, and so on. In a patchily infected forest, the probability of disease in a given tree is dependent on the outcome observed for the previous tree in the transect, and possibly on several previous trees, so there is Markov dependence among trials. Additional factors affect these probabilities, e.g., the distance between trees in the transect, so they also vary over trials.

In animal behavior studies where the categorized behavior is recorded periodically, there is low-order Markov dependence among trials if the time between observations is short or moderate. There is also variation of outcome probabilities across trials, reflecting differences in behavior at different hours of the day, days of the month, and so on.

Radar astronomy. In radar astronomical observations of minor planets or asteroids, data consist of echo power spectral density estimates at a sequence of Doppler frequencies. Empirically determined background filter shape can be removed from the raw spectrum, and the resultant background-free spectrum normalized to the root-mean-square fluctuation in the receiver noise. If no echo is present, the model that the spectral estimates behave as a sequence of IID Normal(0,1) random variables agrees with both *a priori* theoretical considerations and *a posteriori* experimental evidence. However, if the target has a sufficiently large radar cross section, a radar echo would be expected to produce a sequence of above average readings in some portion of the frequency band. A test for the presence of an echo can be based on the length of the longest run of positive readings.

Nonparametric statistics. Tests based on runs, and in particular on the length of the longest run, were mentioned by Gibbons (1971):

> "Since a run which is unusually long reflects a tendency for like objects to cluster and therefore possibly a trend, Mosteller (1941) has suggested a test for randomness based on the length of the longest run."

When a process involves short-term memory or serial dependence, making a model with Markov dependent trials appropriate, assuming the (incorrect) model of IID trials results in very inaccurate run probabilities. Table 1 gives the probability of observing a run of k successes in a sequence of n trials with possible outcomes success (1) and failure (2). Several first- and second-order Markov dependence structures are compared with the IID case. In each case, the transition probabilities are characterized in terms of the theory referred to earlier as "success breeds success". It is clear that an error of only moderate size in specifying the model changes the run probabilities by an order of magnitude, and sometimes by much more.

REFERENCES

Bradley, James V. (1968), Distribution-Free Statistical Tests, Englewood Cliffs, N.J.: Prentice-Hall, Inc.

Feller, William (1968), An Introduction to Probability Theory and Its Applications, Vol. I, 3rd edition, New York: John Wiley and Sons, Inc.

Gibbons, Jean D. (1971), Nonparametric Statistical Inference, New York: McGraw-Hill Book Company.

Mosteller, F. (1941), "Note on an Application of Runs to Quality Control Charts," Annals of Mathematical Statistics, 12, 228-232.

Shaughnessy, Peter W. (1981), "Multiple Runs Distributions: Recurrences and Critical Values," Journal of the American Statistical Association, 76, 732-736.

Table 1. Probabilities of Success Runs of Length k out of n Trials Under Independence and Markov Dependence[a]

Order	Transition Probabilities $p_{11,1}$	$p_{12,1}$	$p_{21,1}$	$p_{22,1}$	Initial Probabilities $p_{00,1}$	$p_{01,1}$	$p_{02,1}$	Type	Probability of Run (k,n) (10,100)	(20,200)	(30,400)
0	.500	.500	.500	.500	.500	.500	.500	IID	.04414	.000087	.00000017
1	.625	.375	.625	.375	.500	.625	.375	SF,M	.22911	.004525	.000084
1	.750	.250	.750	.250	.500	.750	.250	SF,L	.62814	.009363	.011080
1	.625	.500	.625	.500	.500	.625	.500	S,M	.25771	.005166	.000096
1	.750	.500	.750	.500	.500	.750	.500	S,L	.74365	.123016	.014737
2	.750	.500	.500	.250	.500	.625	.375	SF,M	.57888	.083538	.009852
2	.875	.375	.625	.125	.500	.750	.250	SF,L	.90576	.628936	.385788
2	.750	.600	.650	.500	.500	.625	.500	S,M	.72717	.118505	.014183
2	.875	.625	.750	.500	.500	.750	.500	S,L	.99344	.826358	.559748

[a] Type: SF means "success breeds success and failure breeds failure", S means "success breeds success", M means "moderate", L means "large".